Are You *Not* Entertained? Computational Storytelling with Non-Verbal Interaction

Philipp Wicke
philipp.wicke@ucdconnect.ie
University College Dublin
Dublin, Ireland

Tony Veale
tony.veale@ucd.ie
University College Dublin
Dublin, Ireland

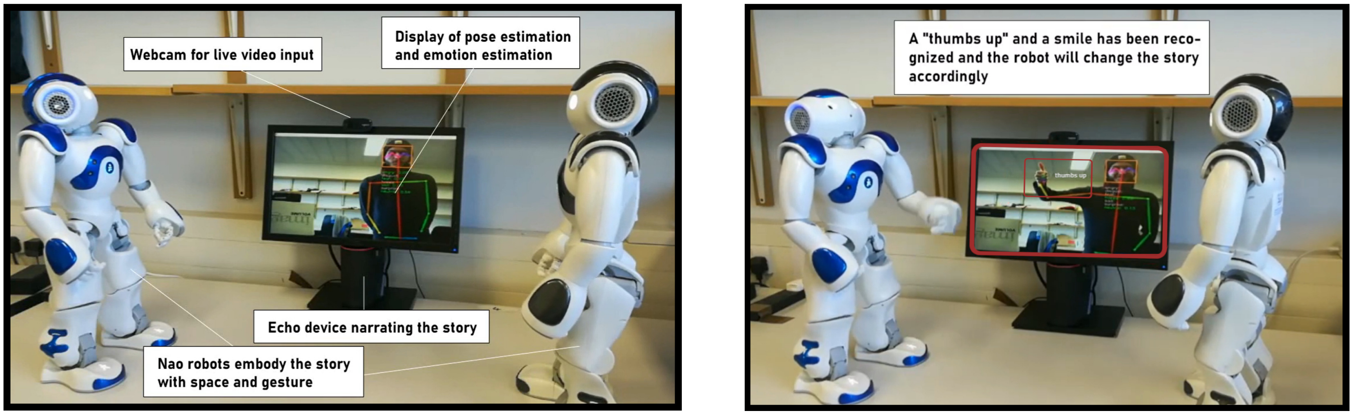
Webcam for live video input
Display of pose estimation and emotion estimation
Echo device narrating the story
Nao robots embody the story with space and gesture
A "thumbs up" and a smile has been recognized and the robot will change the story accordingly
thumbs up

Figure 1: LEFT: Two *Nao* robots act out a story using space and gesture. A display with a webcam provides live-feedback to users while tracking their emotions and gestures. An Echo device narrates the tale. RIGHT: The user provides a thumbs up during storytelling, which alters to course of the story.

ABSTRACT

We describe the design and implementation of a multi-modal storytelling system. Multiple robots narrate and act out an AI-generated story whose plots can be dynamically altered via non-verbal audience feedback. The enactment and interaction focuses on gestures and facial expression, which are embedded in a computational framework that draws on cognitive-linguistic insights to enrich the storytelling experience. With the absence of in-person user studies in this late breaking research, we present the validity of the separate modules of this project and introduce it to the HRI field.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → *Multi-agent systems*; • **Computer systems organization** → *Robotics*.

KEYWORDS

robotics, gestures, storytelling, computational creativity

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3447159

ACM Reference Format:
Philipp Wicke and Tony Veale. 2021. Are You *Not* Entertained? Computational Storytelling with Non-Verbal Interaction. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447159

1 INTRODUCTION

With recent advances in language generation [6, 24], writing stories is no longer left to the creativity of human writers, but shows promising results by machines [5, 42]. However, there is more to storytelling than text generation, and the enactment of a tale through movement and gesture facilitates a greater engagement with the plot [40]. Having an embodied agent, more specifically, a robotic one, can leap a creative AI system from *mere generation* to *true creativity* by providing live-feedback and interaction [36].

Previous work with robotic agents and AI-storytellers made use of minimal interactivity using simple vocal prompts. With prompts like "Yes" and "No", a user was able to pre-determine the plot to be told, by essentially engaging in a co-creative search in a dense graph of branching story lines [38]. In the present implementation, we aim to further engage the user in the telling, and make them feel greater responsibility for plot turns as the tale is told. We look at two modalities, gesture and facial expression, which have previously been studied in the context of storytelling with robots [9, 12, 23], and transfer those modalities onto the interacting human.

In much the same way that Emperor Commodus dictates the fate of the gladiator Maximus in the movie *Gladiator* (see Fig. 2),

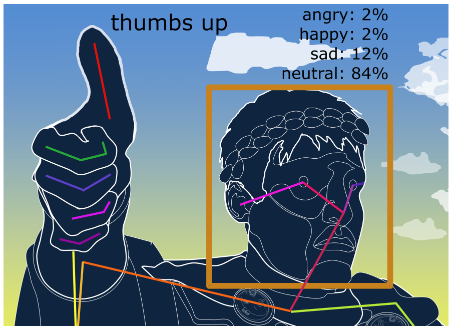

Figure 2: Example of the pose recognition and facial expression recognition software with a cartoon inspired by *Gladiator* (2000) [25]. In this image the emperor decides the fate of Gladiator Maximus with a *verso pollice*, a turned thumb.

we want users to be able to decide the fate of protagonists during a robotic performance. To achieve this, we combine two neural network models: One to classify hand gestures, another to recognize facial emotions. The former identifies diverse hand shapes, but we focus here on the detection of *verso pollice*, the turned thumb, while the latter suggests the sentiment of the action by recognizing emotions on the user's face. As an example, Fig. 2 depicts a cartoon of actor Joaquin Phoenix raising his right arm and showing a thumbs up gesture. An overlay of lines depicts the tracked joints and a boundary box. The latter tracks facial emotions and describes Phoenix as mostly neutral and somewhat sad. This combination of non-verbal signals can prompt the system to alter its story-line in mid-telling, with the robots reacting accordingly. While robotic storytelling is not new [21, 29] and neither are robotic or human gestures for interaction [1, 16], the hallmark of the robot movements used in this work is their schematic foundation. Based on cognitive-linguistic insights from [20, 34, 39], the robots aim to spatially mirror the semantics of each story action. For example, an *insult* action increases emotional distance, so the robots move further apart [41]. When spatial motions and pantomimic gestures are coherently used to mirror the plot, audiences show greater appreciation for a tale and its telling [40]. Gestures by audience members should be equally schematic, even if they exploit a vertical spatial metaphor (up=good, down=bad) rather than a lateral one (closer=positive, farther=negative).

This schematic underpinning is described in greater detail in the system description of the next section, where we outline our approach to its evaluation. The focus of this late-breaking paper is the description and validation of the separate parts of the whole, which are situated in the field of Human-Robot interaction. Furthermore, it justifies the use of gestures for both robots and humans by appealing to a common schematic grounding. The novel contribution of this work, building on past research can be summarized as:

- Remodelling of a pre-existing story generation system to allow live-interaction that alters the course of the story
- Installation of a neural network model in order to allow interaction via gestures
- Installation of a neural network model in order to evaluate the interaction based on facial emotion expression
- Introduction of an approach for spatial and schematic movement in Human-Robot interaction
- Combination of the above into a cohesive system

A holistic evaluation is still pending due to the current impracticality of in-person tests[1].

2 SYSTEM DESCRIPTION

The presented system uses the *Scéalability* storytelling framework [40] in combination with two neural networks: one to recognize hand shapes and another to recognize facial expressions in live images of a human user. After first recapping the relevant previous work, we shall explain how the neural models have been incorporated into the system to permit dynamic interaction with a user, who is presented with the setup shown in Fig. 1. In Fig. 1 two photographic recordings are depicted, both of which show the two robots standing left and right in front of a screen. The screen shows the user with its joints being tracked by a webcam. The image to the left depicts the start of performance with descriptions of the individual devices. The image to the right shows the user on the screen during an interaction where a "thumbs up" and a smile has been tracked by the system. At the start of a performance, the actors (two *Nao* robots) are introduced in character, by a smart speaker (*Amazon Echo*, with *Alexa*[2]). The story-generating system generates a plot and furnishes it with stage directions, narrative and dialogue, and sets up several decision points at which the user may influence the outcome via gestures and facial expressions. Consequently, the robots use their own gestures, spatial movement and speech to act out their parts. For a demonstration of the system in action, readers can watch this online video[3]. The following sections describe all parts of the setup in detail, outline which elements are novel, and explain the research upon which they are built.

2.1 Storytelling Framework

Our storytelling system builds on an existing framework, called *Scéalability* [40], that turns plots into performances by pairing the *Scéalextric* story generator [30, 32] with embodied agents that enact a story using space, gesture and dialogue. The framework has been used to evaluate the effectiveness of different embodiment strategies. In [40], the audience is shown to be sensitive to the coherent use of space in embodied story-telling, where on-stage movements mirror the semantics of plot actions (as opposed to being chosen at random). Although relative spatial movement between actors is more subtle than showy pantomime gestures (as when one robot goes down on one knee to propose), audiences appreciate the former just as much as the latter. The *Scéalextric* story-generator furnishes not only the plot and the dialogue, but also chooses apt characters from a large database of famous figures, real and imaginary, called the NOC list [31].

Notably, *Scéalextric* is pairing figures with respect to character traits and uses well-known tropes, literary stereotypes and a spark of randomness to craft the stories. Due to the vast space of combinations and randomness, some stories can include gender, ethnic or

[1] It was not possible to conduct any in-person research during the COVID-19 pandemic and alternative evaluation strategies are not feasible given the system's setup.
[2] We include *Alexa* to show that the system can also incorporate non-embodied agents
[3] https://youtu.be/xeBgaaOYJXQ

cultural stereotypes. It is a common problem for data-driven story generation systems to learn an inherent bias from the data, which can be addressed with tools such as counterfactual data augmentation [18]. Fortunately, *Scéalextric* is a symbolic AI system, which allows us to track down these stories and label them in order to mitigate bias in the generated content and identify the underlying structures within the database that gives rise to them. Adding mechanisms to our system that identify and reduce undesired biases in the generated stories is important future work.

Two *Nao* robots embody the main characters of the story, while an Echo device narrates the plot. The novelty of the current setup resides in the interaction of story events. Every story can be regarded as a thread of actions. When a decisive action is reached, one in which a character needs to make a decision, the robot will pause, turn towards the audience and ask for a sign as to what it should decide. While the robot speaks directly to the audience, a user can provide a thumbs up or thumbs down (just like Emperor Commodus), with a corresponding facial expression. Users need not react at all, but if they do, their inputs decide the robot's next action and the plot is regenerated from that point. The evaluation of user responses is explained in Section 2.5.

The underlying *Scéalextric* system has been adapted so that it can account for whatever decision the user makes. If the decision is in-line with the previously generated plot, the narrative continues as planned. If it is different, *Scéalextric* suggests a *detour* that sets the story on a different path. As shown in [33], each story is mostly self-correcting. When user decisions force a detour, the plot still rejoins its original narrative arc before the conclusion is reached.

2.2 Robotic Enactment

This focus on gestures (by actors and users) and on-stage movements (by actors) allows us to incorporate aspects of cognitive-linguistic theory into the system. This is a small but novel application of theory to what is otherwise a practical AI system.

2.2.1 *Spatial/Schematic Movement.*

A common language for interaction between two or more embodied systems is not necessarily a spoken one. Rather, embodied agents can make use of recurring structures, called *image schemas* [15], that are shaped by bodily experience of a shared environment [19]. For example, our experience of gravity makes vertical orientation, of something being "up" or "down," a salient basis for human metaphors and thus for machine sensors too. Evidence for these schemas are pervasive in natural language, and provide a combinatorial basis for defining a computational semantics [3, 14, 17]. Moreover, this spatial framework supports a practical formalism for robotic movement [4, 28]. Image schemas are not abstract, since they are grounded in physical experience, but they are generic, and underpin creative tasks such as conceptual blending [13] and storytelling [39], and provide a conceptual basis for human gestures [8, 20]. Using robots that use both spatial movements and gestures to accentuate meaning, [40] show how certain on-stage actions can concisely summarize the cumulative state the narrative. For instance, in a tale that brings two characters closer together emotionally, the robots will move physically closer over time [41]. In this way, their relative position at any one time sums up the story so far, This universal metaphor helps to explain why audiences appreciate these movements as much as pantomimic gestures which are transitory and non-cumulative.

The robot actors in [41] do not exploit the vertical dimension, either literally or metaphorically. However, this dimension is also grounded in cognitive image schemas, and we exploit it here to support language-free human intervention into a story.

2.2.2 *Other modalities.*

Besides moving back and forth on stage, the robots also perform pantomimic movements (bowing, waving, kneeling, pointing, etc.), speak their lines via a speech synthesizer, and coordinate their actions with the Echo device that narrates the action. Specifically, we use bipedal anthropomorphic *Nao* robots from *Softbank Robotics* with 25 degrees of freedom, four microphones, two HD cameras and a variety of sensors for detecting pressure, inertia and infrared light [11]. The pantomimic movements that punctuate the events of the story are adopted from [37, 38], who explain how more than 400 story verbs are mapped onto a corresponding number of gestures via a simple probabilistic model. The validity of this mapping has been demonstrated through empirical evaluation of robotic story performances [40].

The *Amazon* Echo device that narrates the story has been chosen for its clear audio output. It has been used in combination with *Nao* robots for related creative tasks [35] and is easily integrated via *Scéalability.* Since the *Nao*'s speech synthesis falls short of *Alexa*'s, Google's *WaveNet* is used for Text-To-Speech synthesis in real-time [22]. *Scéalability* provides information about the gender of each character, and this is used to inflect the spoken output accordingly.

2.3 Hand Shape Recognition

Our system tracks the user's pose in real time using the *OpenPose* framework for *Python* [7][4], which provides 135 keypoints on single images of multiple joint positions (torso, head, legs, arms, hands and face). The keypoints of the hands [27] are the inputs to a neural network that is trained to classify different hand shapes[5]. The model classifies nine different hand shapes (with an accuracy of 96%), of which we seek only the thumbs up and down gestures. Although gestures are culturally dependent, the thumbs up gesture is commonly understood as "good" or "positive" across cultures [26]. Figure 2 shows how *OpenPose* provides a colored skeleton of the limbs, head and fingers, even when parts of the arm are occluded. The hand shape model classifies the finger and displays the label on screen ("Thumbs up" in the example).

Whenever a thumbs up or down is detected in the camera-input, a marker is recorded for the next decision point in the storytelling process. The lack of a hand signal is deemed a neutral response. The blackboard architecture of the *Scéalability* framework, which allows backstage coordination between the robot actors and the Echo device, also allows the cameras and classifiers to be integrated into the system as yet another information source for storytelling.

2.4 Facial Expression Recognition

An additional sentiment accompanying the hand shape is provided by facial expression, so that an angry thumbs up carries a different meaning than a surprised one. We use the *Python* package *FER*[6] for

[4] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[5] https://github.com/Fasko/Hand-Gesture-Recognition
[6] https://github.com/justinshenk/fer

Facial Expression Recognition [10]. This identifies faces with the MTCNN face detector [43] and uses a convolutional network to classify expressions [2] by emotion, whether *Angry, Happy, Sad, Surprise, Fear, Disgust* or *Neutral.* Arriaga, Plöger and Valdenegro (2017) report human-level performance for their network architecture.

Figure 2 shows the boundary box for the face of emperor Commodus, as drawn by the MTCNN face detector, while *FER* displays percentages for the most salient emotions. We set a confidence threshold of 80% in order to reliably identify the most salient expressions for audience feedback: *Angry, Happy* and *Surprise.* These three prove to be the most robust classes in our setup. We then combine this information with classified hand shapes to derive a meaningful signal from the user to guide the story.

2.5 Signal Evaluation and Enactment

The detected hand signals (thumbs up, thumbs down, no hand) and facial expressions (happy, angry, surprise and neutral) can combined to twelve (3x4) combinations such as "*enthusiastic thumbs up*" when a thumbs up and a happy smile is detected. The robot re-articulates the combination that it detects and alters the plot if a narrative detour is necessitated. For example, the robot in the linked video asks "*Should I give my heart to this person or not?*" When the user provides a thumbs up and smiles, this is interpreted as an *enthusiastic thumbs up*, and the narrative is adapted accordingly. Each gesture is interpreted in the context of the robot's specific request, so a thumbs down in response to the question "Should I turn down this offer" is interpreted as a yes (turn it down), while a thumbs up to the suggestion of a marriage proposal is also understand as a yes (do propose). A knowledge-base connects signals and event actions to appropriate decisions. Examples of decisive actions and the signal valences (summed for hands and face) that motivate them are provided in Table 1 for responses to robot questions of the form "*Shall I ...*". If the robot asks "*Shall I lay a trap for ...* ", a frown is not necessarily negative, but can express Schadenfreude about the deed. Likewise,the thumbs down has a positive valence for the question "*Shall I kill ...* " since it puts the user in the same position as emperor Commodus. Given this context-specificity, the signal-interpretation matrix requires fine-tuning and evaluation through additional studies and experiments.

3 FUTURE WORK

Upon the resumption of in-person experimentation, an empirical evaluation will be conducted to provide a perspective on the perception of the system. We plan to have users interact with the system in three versions: First, the system will ask the audience for feedback, but their response will not affect the story (interaction without influence). Second, the system will ask the audience for verbal-only feedback (yes-no interaction). Third, the system would work as described in this paper. We plan to adopt the methodology presented in [40]. Our three conditions will allow us to tease apart the effect of the interaction itself and the effects of the non-verbal interaction. Additionally, we plan to use the results from this study to fine-tune our signal encoding (Table 1).

Considering the issues about inappropriate bias in generated stories, we work towards implementing mechanisms in the proposed system to identify and prevent generating storylines that can be perceived as offensive, such as corpus-level constraints [44].

Should I ...		👍	👎	🙂	😠	😮	😐
fall in love with	0.0	+0.9	-0.9	+0.4	-0.4	-0.3	0.0
propose to	0.0	+0.9	-0.9	+0.4	-0.4	0.0	0.0
lay a trap for	0.0	+0.9	-0.9	+0.4	0.0	-0.4	0.0
rise against	0.0	+0.9	-0.9	+0.4	+0.4	-0.4	0.0
rebel against	0.0	+0.9	-0.9	+0.4	+0.4	-0.4	0.0
stand up to	0.0	+0.9	-0.9	+0.4	+0.4	-0.4	0.0
turn against	0.0	-0.9	+0.9	+0.4	-0.4	-0.5	0.0
...	...	...	...	...	...	...	...

Table 1: Signal encoding examples of hand gestures and facial expressions for decisive questions during storytelling.

4 CONCLUSION

This short, late-breaking paper sketches the architecture of a interactive robotic storytelling system that integrates non-verbal cues from both the performers (robots) and the audience (human users). Our focus is on the naturalness of these cues, which are grounded in pantomime (iconic actions that are the gestural equivalent of idioms and cliches) and spatial metaphor (in particular, schematic movement that has an emotional interpretation). An evaluation of the relative merits of pantomime and metaphor for robot performance has already been conducted, and reported in summary here; subsequent evaluations of the effectiveness of physical feedback from the audience are next on our agenda.

The use of plot-driven gestures and spatial movements have already been evaluated and shown to be appreciated by audiences [40], while the usefulness of gestural inputs and facial expressions for HRI have also been evaluated elsewhere [9, 12, 23].

The novelty of this approach to HRI lies in its heterogeneous combination of cognitive linguistic models of space, symbolic AI approaches to story generation, robotic models of performance, and neural network models of visual signal detection. This combination is expedient but far from shallow, since each element must ultimately connect at the plot level [32]. Gestures and movements and user signals must all integrate with an explicit sense of what is happening in the tale, which the system represents at both a surface text level (narration and dialogue) and a deep semantic level.

In an important sense, then, the system benefits from its heterogeneity. Although neural language models generate fluent textual narratives [6, 24], the plot-driven use of space and gesture necessitates access to the symbolic deep-structures of the narrative that are not apparent at the surface level. The authors eagerly await the resumption of social interaction and in-person experimentation to further evaluate and improve this assemblage through feedback from the end-users and the HRI community more generally.

ACKNOWLEDGMENTS

The authors would like to thank Michael Fasko Jr and Jacob Calfee for access to the hand shape recognition model, Justin Shenk for the implementation of *FER* and the HRI21 LBR Co-Chairs and reviewers.

REFERENCES

[1] Dimitra Anastasiou, Kristiina Jokinen, and Graham Wilcock. 2013. Evaluation of WikiTalk–user studies of human-robot interaction. In *International Conference on Human-Computer Interaction*. Springer, 32–42.

[2] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557* (2017).

[3] John A Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence* 174, 14 (2010), 1027–1071.

[4] Daniel Beßler, Robert Porzel, Mihai Pomarlan, Michael Beetz, Rainer Malaka, and John Bateman. 2020. A Formal Model of Affordances for Flexible Robotic Task Execution. In *ECAI. Proc. of the 24th European Conference on Artificial Intelligence.*

[5] Gwern Branwen. 2020. GPT-3 Creative Fiction. (2020). https://www.gwern.net/GPT-3#poetry

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[8] Alan Cienki. 2005. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 421–442.

[9] Rodolphe Gelin, Christophe d’Alessandro, Quoc Anh Le, Olivier Deroo, David Doukhan, Jean-Claude Martin, Catherine Pelachaud, Albert Rilliard, and Sophie Rosset. 2010. Towards a storytelling humanoid robot. In *2010 AAAI Fall Symposium Series*.

[10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*. Springer, 117–124.

[11] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. 2009. Mechatronic design of NAO humanoid. In *2009 IEEE International Conference on Robotics and Automation*. IEEE, 769–774.

[12] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan. 2011. Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In *International conference on social robotics*. Springer, 71–83.

[13] Maria M Hedblom, Oliver Kutz, and Fabian Neuhaus. 2016. Image schemas in computational conceptual blending. *Cognitive Systems Research* 39 (2016), 42–57.

[14] Maria M Hedblom, Oliver Kutz, Rafael Peñaloza, and Giancarlo Guizzardi. 2019. Image schema combinations and complex events. *KI-Künstliche Intelligenz* 33, 3 (2019), 279–291.

[15] Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

[16] Jhonatan Kobylarz, Jordan J Bird, Diego R Faria, Eduardo Parente Ribeiro, and Anikó Ekárt. 2020. Thumbs up, thumbs down: non-verbal human-robot interaction through real-time EMG classification via inductive and supervised transductive transfer learning. *Journal of Ambient Intelligence and Humanized Computing* (2020), 1–11.

[17] Werner Kuhn. 2007. An image-schematic account of spatial categories. In *International Conference on Spatial Information Theory*. Springer, 152–168.

[18] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, 189–202.

[19] Jean M Mandler. 1992. How to build a baby: II. Conceptual primitives. *Psychological review* 99, 4 (1992), 587.

[20] Irene Mittelberg. 2018. Gestures as image schemas and force gestalts: A dynamic systems approach augmented with motion-capture data analyses. *Cognitive Semiotics* 11, 1 (2018).

[21] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 518–523.

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[23] Catherine Pelachaud, Rodolphe Gelin, Jean-Claude Martin, and Q Anh Le. 2010. Expressive gestures displayed by a humanoid robot during a storytelling application. *New frontiers in human-robot interaction (AISB), Leicester, GB* (2010).

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[25] Scott Ridley. 2000. *Gladiator*. DreamWorks Pictures.

[26] Joel Sherzer. 1991. The Brazilian thumbs-up gesture. *Journal of Linguistic Anthropology* 1, 2 (1991), 189–197.

[27] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[28] Michael Spranger and Martin Loetzsch. 2009. The semantics of SIT, STAND, and LIE embodied in robots. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society (Cogsci09)*. Cognitive Science Society, 2546–2552.

[29] Hendrik Striepe, Melissa Donnermann, Martina Lein, and Birgit Lugrin. 2019. Modeling and Evaluating Emotion, Contextual Head Movement and Voices for a Social Robot Storyteller. *International Journal of Social Robotics* (2019), 1–17.

[30] Tony Veale. 2016. A Rap on the Knuckles and a Twist in the Tale From Tweeting Affective Metaphors to Generating Stories with a Moral. In *2016 AAAI Spring Symposium Series*.

[31] Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. 34–41.

[32] Tony Veale. 2017. Déjà Vu All Over Again: On the Creative Value of Familiar Elements in the Telling of Original Tales.. In *ICCC*. 245–252.

[33] Tony Veale. 2018. Appointment in Samarra: Pre-destination and Bi-camerality in Lightweight Story-Telling Systems.. In *ICCC*. 128–135.

[34] Tony Veale and Mark T Keane. 1992. Conceptual Scaffolding: A spatially founded meaning representation for metaphor comprehension. *Computational Intelligence* 8, 3 (1992), 494–519.

[35] Tony Veale, Philipp Wicke, and Thomas Mildner. 2019. Duets Ex Machina: On The Performative Aspects of" Double Acts" in Computational Creativity. In *ICCC*. 57–64.

[36] Dan Ventura. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the Seventh International Conference on Computational Creativity*. Sony CSL, Paris, 17–24.

[37] Philipp Wicke and Tony Veale. 2018. Interview with the Robot: Question-Guided Collaboration in a Storytelling System. In *Proc. of ICCC’18, the International Conf. on Computational Creativity*. 56–63.

[38] Philipp Wicke and Tony Veale. 2018. Storytelling by a Show of Hands: A framework for interactive embodied storytelling in robotic agents. In *Proc. of AISB’18, the Conf. on Artificial Intelligence and Simulated Behaviour*. 49–56.

[39] Philipp Wicke and Tony Veale. 2018. Wheels Within Wheels: A Causal Treatment of Image Schemas in An Embodied Storytelling System.. In *TriCoLore (C3GI/ISD/SCORE)*.

[40] Philipp Wicke and Tony Veale. 2020. The Show Must Go On: On the Use of Embodiment, Space and Gesture in Computational Storytelling. *New Generation Computing* (2020), 1–28.

[41] Philipp Wicke and Tony Veale. 2020. Walk the Line: Digital Storytelling as Embodied Spatial Performance. In *7th Computational Creativity Symposium at AISB 2020*.

[42] Alexey Tikhonov Yana Agafonova and Ivan P. Yamshchikov. 2020. Paranoid Transformer: Reading Narrative of Madness as Computational Approach to Creativity. In *Proceedings of the 11th International Conference on Computational Creativity*. Association for Computational Creativity, Coimbra, Portugal, 146–152. http://computationalcreativity.net/iccc20/papers/037-iccc20.pdf

[43] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[44] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989.